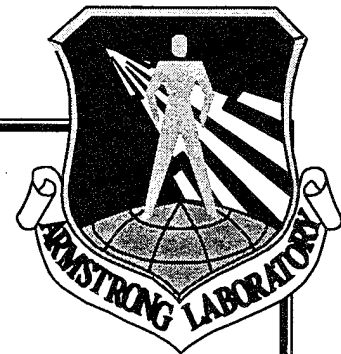AL/HR-TR-1995-0092

ARMSTRONG LABORATORY

SOFTWARE AND PROGRAMS FOR CONDUCTING META-ANALYSIS RESEARCH: A MONTE CARLO INVESTIGATION OF POTENTIAL DIFFERENCES

Winfred Arthur, Jr.

Department of Psychology
Texas A&M University
College Station, Texas 77843-4235

Winston Bennett, Jr.

HUMAN RESOURCES DIRECTORATE
TECHNICAL TRAINING RESEARCH DIVISION
7909 Lindbergh Drive
Brooks AFB, Texas 78235-5352

Allen I. Huffcutt

Department of Psychology
Bradley University
Peoria, Illinois 61625

19960130 041

**AIR FORCE MATERIEL COMMAND
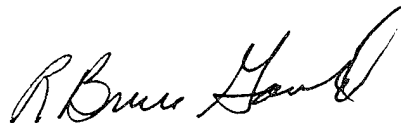BROOKS AIR FORCE BASE, TEXAS**

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.
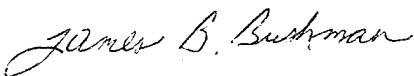
This paper has been reviewed and is approved for publication.

WINSTON BENNETT, JR
Project Scientist
Instructional Systems Research Branch

R. BRUCE GOULD, Technical Director
Technical Training Research Division

JAMES B. BUSHMAN, LtCol, USAF
Chief, Technical Training Research Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE June 1995	3. REPORT TYPE AND DATES COVERED Interim - Sep 1991-Dec 1994

4. TITLE AND SUBTITLE

Software and Programs for Conducting Meta-Analysis Research: A Monte Carlo Investigation of Potential Differences

5. FUNDING NUMBERS

PE-62205F
PR-1121
TA-12
WU-00

6. AUTHOR(S)
Winfred Arthur, Jr.   Allen I. Huffcut
Winston Bennett, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Texas A&M University   Bradley University
Department of Psychology   Department of Psychology
College Station, TX 77843-4235   Peoria, IL 61625

8. PERFORMINGORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Armstrong Laboratory
Human Resources Directorate
Technical Training Research Division
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AL/HR-TR-1995-0092

11. SUPPLEMENTARY NOTES
Armstrong Laboratory Technical Monitor: Winston Bennett, Jr. (210) 536-2932; DSN: 240-2932

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

With the increased popularity of meta-analysis, a number of articles have questioned and highlighted the role that judgment calls play in the implementation and, ultimately, outcomes of studies that use this technique. In the absence of standardized data analyses packages, coupled with the wide range of options available to meta-analysts, the current study investigated the effect of choice of data analysis programs on meta-analysis study outcomes. The objective of this Monte Carlo study was to investigate the extent to which four commonly used Schmidt and Hunter validity generalization-based meta-analysis software programs, all based on the same conceptual and theoretical assumptions, produced identical outcomes when used to analyze the same dataset. The results indicate that while there were some differences in values obtained from the programs, these differences tended to be very small, typically occurring in the fourth and sometimes fifth decimal place, and did not influence the meta-analytic *outcomes*. Finally, differences in the features and capabilities of each of the programs are presented and discussed.

14. SUBJECT TERMS		15. NUMBER OF PAGES
Meta-Analysis Software Programs Monte Carlo Data Validity Generalization Quantitative Review		34
		16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. Z39-18
298-102

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

With the increased popularity of meta-analysis, a number of articles have questioned and highlighted the role that judgment calls play in the implementation and, ultimately, outcomes of studies that use this technique. In the absence of standardized data analyses packages, coupled with the wide range of options available to meta-analysts, the current study investigated the effect of choice of data analysis programs on meta-analysis study outcomes.

The objective of this Monte Carlo study was to investigate the extent to which four commonly used Schmidt and Hunter validity generalization-based meta-analysis software programs, all based on the same conceptual and theoretical assumptions, produced identical outcomes when used to analyze the same dataset. The results indicate that while there were some differences in values obtained from the programs, these differences tended to be very small, typically occurring in the fourth and sometimes fifth decimal place, and did not influence the meta-analytic *outcomes*. Finally, differences in the features and capabilities of each of the programs are presented and discussed.

An earlier version of this paper was presented at the Seventh annual convention of the Society for Industrial and Organizational Psychology, Montreal Canada, April 1992.

Software and Programs for Conducting Meta-Analysis Research:

A Monte Carlo Investigation of Potential Differences

SUMMARY

This research examined the extent to which four common meta-analytic programs and software, which were all based on the same conceptual and theoretical assumptions, produced identical outcomes when used to analyze a common dataset. The results indicated some differences in outcome values obtained from the programs, however these differences tended to be very small, and did not influence the meta-analytic *outcomes*. There were substantive differences in the statistics provided by each program, and differences in the features and capabilities of each of the programs. Implications for future meta-analysis research and the differences found between the programs used in this study are presented and discussed.

I. INTRODUCTION

Meta-analysis is a statistical tool for summarizing empirical results across a number of studies to reach a quantitative generalization. While there are a number of meta-analytic approaches and techniques, Hunter and Schmidt's validity generalization procedure (Hunter & Schmidt, 1990) is the most commonly used meta-analysis technique in organizational behavior/human resource management (OB/HRM) (Steiner, Lane, Dobbins, Schnur, & McConnell, 1991). Therefore this paper is limited to an examination of the validity generalization approach to meta-analysis.

Background

Meta-analysis. One reason for the increasing popularity of meta-analysis in OB/HRM, and in social science in general (Wanous, Sullivan, & Malinak, 1989) is that, compared to alternative methods, it is a quantitative and relatively more objective method for summarizing the empirical literature in a given domain (Hunter & Schmidt, 1990). Paradoxically, coupled with its increased popularity are a number of articles that have questioned and highlighted the

role judgment calls play in the *implementation* (and ultimately, outcomes) of meta-analyses (e.g., Wanous et al., 1989).

Judgment Calls. The effect of judgment calls in the implementation of meta-analysis have been shown to account for differences in ostensibly objective meta-analytic studies of the same content area (see Wanous et al., 1989 for a review of some of these studies). Specifically, the following seven steps in the implementation sequence have been noted to call for some judgment on the part of the meta-analyst (e.g., Wanous et al., 1989): (1) topic selection--defining the research domain; (2) specifying the inclusion criteria; (3) searching for and locating relevant studies; (4) sampling and selecting the final set of studies; (5) extracting data and coding study characteristics; (6) deciding to group or separate multiple measures of independent and dependent variables; and (7) selecting potential moderators.

Problem

Given the relatively recent emergence of meta-analysis techniques and the unavailability of *standardized* procedures to conduct meta-analyses on readily available statistical software packages such as SAS, SPSS, and BMDP, the *data analysis* step in the implementation sequence (i.e., calculating mean correlations and correcting for artifacts) could *also* be considered and investigated as one that calls for some judgment and a decision on the part of the researcher. Consequently, there are several reasons for this investigation. First, unlike other widely used statistical techniques such as *t* tests, analysis of variance, and measures of association (e.g., Pearson's *r*), which are readily available on the previously mentioned statistical software packages, the researcher has to make a decision as to how the data analysis (i.e., calculating mean correlations and correcting for artifacts) will proceed when conducting a meta-analysis. These choices range from using one of several available programs without

any modifications, modifying these programs, writing one's own program based on the available correction formulas, to maybe even doing some or all the calculations by hand. Reflecting this is the observation that published meta-analyses have used a variety of software programs for their data analyses. For example, McDaniel, Schmidt, and Hunter (1988) used a program developed by McDaniel (1986a, 1986b) which is based upon procedures outlined by Hunter and Schmidt. This program consists of a series of SAS macros designed to calculate a variety of meta-analysis statistics. Lord, DeVader, and Alliger (1986) and Arthur, Barrett, and Alexander (1991) used a FORTRAN program originally developed by Schmidt and associates. Barrick and Mount (1991) used Hunter and Schmidt's (1990) BASIC meta-analysis program for microcomputers. Unfortunately, in addition to being "unstandardized," there is as yet no published data on the similarity or equivalence of the outcomes of these programs. Furthermore, there are also other studies that discuss meta-analysis procedures but do not indicate what program or software, if any, was used for the data analysis (e.g., Schmitt, Gooding, Noe, & Kirsch, 1984; Wiesner & Cronshaw, 1988).

Although there may be differences in outputs of commercially available statistical packages, such as SAS and SPSS, for a given statistic (e.g., t or F) computed on the same dataset, these differences are typically manifested in the fourth decimal place or higher. Thus, in reading a study that uses a t test to report differences between two means, there is little reason to question the accuracy of the software that was used to run the analysis. The same level of confidence may not hold true for meta-analysis studies. As noted earlier, with several researchers writing their own meta-analysis programs (e.g., Arthur et al., 1991; Huffcutt, Arthur, & Bennett, 1993; Hunter & Schmidt, 1990; McDaniel, 1986a, 1986b), it is not untenable to speculate that although conceptually similar, different meta-analysis programs

3

could result in non-identical results. One plausible reason for this has to do with potential differences in the formulas used to calculate the various statistics and corrections for artifacts. A second and equally plausible reason could simply be typographical and computational errors in creating and writing the programs. Ironically, typographical and computational errors are listed by Schmidt and Hunter as one source of artifactual error, and as noted by Hunter and Hirsh (1987), "these errors can be very large in magnitude" (p. 322).

Thus, given the wide spread use of meta-analytic procedures, the absence of any standardized data analysis packages, and the failure of most published studies to indicate what program or software was used for the data analysis, it seemed important to investigate whether the choice of meta-analytic data analysis program would influence study outcomes. As previously indicated, one criticism leveled at meta-analysis is that while ostensibly objective, there are typically several judgments that must be made in the implementation of a meta-analytic study. With the availability of several programs, the choice of which meta-analysis program to use becomes *yet* another judgment call to be considered. Of course, if the programs produce identical outcomes then the effect of this judgment is minimized. But to the extent that they do not, this becomes a critical judgmental step in the meta-analytic study process.

Study Objective

The objective of the current study was to investigate the extent to which different meta-analytic programs and software, that are based on the same conceptual and theoretical assumptions, produced the same or different outcomes when used to analyze the same dataset. The exploratory nature of the present study precluded the specification of any hypothesis. A Monte Carlo dataset with a specified true validity was generated. Predictor and criterion

unreliability and range restriction were then induced to attenuate the correlations. Four meta-analysis programs were next used to analyze the attenuated ("observed") correlations to assess the extent to which they would differ in correcting back to the true validity.

Validity Generalization Software and Programs. The choice of software programs was limited to those based on the Schmidt and Hunter validity generalization procedures. There were two reasons for this decision. First, as noted by Steiner et al. (1991), the Schmidt and Hunter validity generalization procedure is the most commonly used meta-analysis procedure in OB/HRM, thus it seems to be widely accepted as the procedure of choice. Second, limiting the programs to only those based on Schmidt and Hunter's validity generalization procedures ensured that they were based on the same conceptual and theoretical assumptions and consequently should be analyzing the data in the same fashion. After an exhaustive search and review of the literature and consultation with some researchers and colleagues engaged in meta-analytic research, four programs that met the above criteria were identified. These were (1) a FORTRAN program originally developed by Schmidt and associates; (2) Huffcutt et al.'s (1993) SAS PROC MEANS program; (3) Hunter and Schmidt's (1990) BASIC program for microcomputers; and (4) McDaniel's (1986a) MAME program. The dataset and programs are described in detail in the method section.

II. METHOD

Dataset

To generate a realistic Monte Carlo dataset, the statistical properties of the data were set to be similar to the characteristics of ability tests for entry-level jobs reported in Hunter and Hunter's (1984) meta-analysis. Specifically, as reported in Hunter and Hunter (1984), the dataset consisted of 32,124 subjects distributed across 425 studies. The average number of subjects per study was 75.586 ($SD = 25$) and for the population of 32,124 subjects, the correlation between predictor and criterion scores was set at .53. The primary advantage to using such a customized dataset is that it provides known starting values to be used as true scores against which to compare the outputs of the different programs.

True r's. The first step in building the dataset called for developing a randomly generated predictor score distribution with a mean of 100 and a standard deviation of 15 for 32,124 "subjects." These descriptive statistic parameters were chosen to be consistent with those typically reported for general ability tests. Next, a criterion score was randomly assigned to each subject such that the population ($N = 32,124$) validity coefficient was .53. Accordingly, .53 represented the "true" population correlation. Subjects were next randomly assigned to 425 studies such that the average number of subjects per study was 75.586 ($SD = 25$). Predictor/criterion correlations were then computed for each study. These coefficients are subsequently referred to as "true correlations."

Levels of Reliability. Next, using procedures consistent with other Monte Carlo validity generalization studies (e.g., Callender & Osburn, 1980), the true correlations were attenuated by inducing predictor and criterion unreliability, and range restriction; these were induced simultaneously. To accomplish this, a reliability factor was first calculated for each of

the 425 studies. Predictor reliabilities of .75, .80, .85, .90, and .95 were selected. These levels are consistent with those reported for commonly used general ability tests. The 425 studies were randomly assigned to the five reliability levels such that 85 studies were assigned to each level. Simulating supervisor ratings, a criterion reliability of .60 (Rothstein, 1990) was chosen. The reliability factor for each of the 425 study correlations was calculated using the following formula (Ghiselli, Campbell, & Zedeck, 1981):

$$REL\_F = (\sqrt{r_{xx}}) (\sqrt{r_{yy}}) \tag{1}$$

where *REL_F* is the reliability factor (i.e., predictor and criterion unreliability); r_{xx} is the predictor reliability (.75, .80, .85, .90, or .95); and r_{yy} is the criterion reliability (.60).

Range Restriction. Next, a range restriction factor was calculated. To remain consistent with our simulation of the Hunter and Hunter (1984) dataset, the range restriction ratio (*u*) for each of the 425 studies was assumed to be .67 (Hunter & Hunter, 1984, p. 79). The range restriction factor for each study was calculated using the following formula (Hunter & Schmidt, 1990, p. 48):

$$RR\_F = u / [(u^2-1) \rho^2 + 1]^{\frac{1}{2}} \tag{2}$$

where *RR_F* is the range restriction factor; *u* is the range restriction ratio (.67); and ρ is the population correlation (.53).

Attenuation. The final step in creating the Monte Carlo dataset was to induce both unreliability and range restriction for each of the 425 study correlations. This was accomplished by multiplying each study correlation by a reduction factor using the formula below:

$$Obs\ r_{xy} = (True\ r_{xy}) (REDUC\_F) \tag{3}$$

7

where *Obs r_{xy}* is the attenuated (observed) correlation (i.e., predictor and criterion unreliability, and range restriction induced); *True r_{xy}* is the unattenuated correlation; *REDUC_F* is the product of the reliability and range restriction factors (i.e., *REL_F * RR_F*).

So, having started off with known "population" values and then inducing measurement attenuation and range restriction (artifacts), the ultimate test of the accuracy of the meta-analysis programs was to assess the extent to which they would correct *back* to the known population values. Specifically, the programs, if accurate, should return to a true mean corrected *r* of .53 (*SD* = 0) after correcting the observed correlations for measurement attenuation and range restriction.

Programs

As previously indicated, the choice of software programs was limited to those based on the Schmidt and Hunter validity generalization procedures. Software copies of the programs were obtained from their developers or their associates.

The FORTRAN Program. This program was designed to run on main frame computers. It was originally developed by Schmidt and associates, but the version used here has since been modified at the University of Akron to compute additional statistics such as a 90% and 95% confidence interval about the mean *r* and sample size. Although this program corrected for sampling error only, it was, nevertheless, used in the analysis because as noted by Steiner et al. (1991), of the studies using the Hunter and Schmidt procedure, 100% controlled for sampling error. This is in contrast to only 69.2% for unreliability in the criterion, 50% for unreliability in the predictor, and 11.5% for range restriction in the predictor. In addition to reporting a variety of meta-analysis statistics, the program also reports statistics computed on

both the raw data (i.e., r's) and the Fisher z transformation. This program has been used in several published studies including Arthur et al. (1991) and Lord et al. (1986).

Huffcutt, Arthur, and Bennett's (1993) SAS PROC MEANS Program. This program uses the PROC MEANS procedure in SAS to compute most of the calculations and summary statistics called for in the Schmidt and Hunter meta-analysis approach. The program uses the noninteractive approach. Like the Hunter and Schmidt and McDaniel programs used in the current study, this program uses the artifact distribution approach. Specifically, it analyzes correlations using artifact distributions. A limitation of this program is that even though relatively simple and easy, a few of the final calculations have to be made manually. This program has been used in a number of studies including Huffcutt and Arthur (1994). A detailed description of this program can be found in Huffcutt et al. (1993).

Hunter and Schmidt's (1990) VGNON.BAS BASIC Program. Of the four programs used in this paper, this is the only one that runs specifically on microcomputers. Hunter and Schmidt (1990) presented four GW-BASIC meta-analysis programs. However, to remain consistent with the other programs used in the study, the program (VGNON.BAS), which analyzes correlations using artifact distributions, was chosen. This is a noninteractive program that corrects for predictor and criterion unreliability and range restriction, and has been used in a number of studies including Barrick and Mount's (1991) meta-analysis of the relationship between personality and job performance.

McDaniel's (1986a) MAME Program. This program consists of a series of SAS macros designed to calculate a variety of meta-analysis statistics. Like Hunter and Schmidt (1990), McDaniel (1986a) provided several meta-analysis programs. However, for this investigation, a macro (%RMETA) was chosen along with options that used artifact distributions to analyze

the correlations. The program uses the interactive approach and corrects for predictor and criterion unreliability and range restriction. It has been used in several meta-analytic studies including McDaniel et al. (1988a).

Data Analysis

The 425 attenuated correlation coefficients in conjunction with the artifact distributions were analyzed using the four meta-analysis programs described in the previous section. The artifact distributions used to analyze the correlations were the same as those used to induce measurement attenuation. Specifically, these were criterion unreliability of .60 (frequency = 425), and predictor reliabilities of .75, .80, .85, .90, and .95 (each with a frequency of 85 respectively). The range restriction ratio (u) for each study was set to .67.

The FORTRAN, Huffcutt et al. (1993), and McDaniel (1986a) programs were all run on a mainframe computer. With the exception of the variable input statements, no other changes or alterations were made to the program codes. Some job control language (JCL) changes, required by the local system, had to be made to the McDaniel (1986a) and FORTRAN programs before they could be run. It must be emphasized that these changes were not to the program code but were limited to the JCL.

The Hunter and Schmidt (1990) program was run on a microcomputer. It was first converted from GW-BASIC to QuickBASIC 4.50 and then run in this environment (because the authors were more proficient with the latter than the former). One programming change had to be made to the Hunter and Schmidt (1990) program. The copy obtained by the authors, along with that presented in Hunter and Schmidt (1990) limits the number of correlations that can be analyzed to 100 cases only (i.e., a BASIC dimension array statement had been set to

100). Because 425 correlations were being analyzed, this problem was rectified by changing the value of the dimension array (statement) to 425.

III. RESULTS

Results of the comparative analysis are presented in Table 1. The reported meta-analysis statistics have been limited to those common to at least two programs.

Table 1
Comparative Meta-Analysis Statistics for the Four Programs

Meta-Analysis Statistic	True Value	Programs			
		FORTRAN Program	Huffcutt et al. (1993)	Hunter and Schmidt (1990)	McDaniel (1986a)
SAMPLING ERROR RESULTS					
Total N	32,124	32,124	32,124	32,124	32,124
Number of r's	425	425	425	425	425
Mean Observed r (sample-weighted)	[0.27605][A]	0.27632 **0.27721**	0.27638	0.27636	0.27636
SD Observed r's	[0.05845][A]	0.05366 **0.05818**	0.05360	0.05360	0.05360
Var Due Sampling Error	--	--	0.01143	0.01139	0.01139
Percent Var Accounted For	--	395.75537 **410.41723**	398.016	396.7655	396.76548
Residual Var	--	--	-0.00856	--	-0.00852
Chi Square	--	108.45719 **103.36209**	--	--	107.11617
ATTENUATION ARTIFACT CORRECTIONS					
Mean True r	0.53	--	0.55283	0.55276	0.55276
SD of True r	0.09452	--	0.00	0.00	0.00
Percent Var Due to All Artifacts	--	--	402.640	408.9199	401.38798
Residual Var	--	--	-0.08695	-0.00875	-0.00865
MEAN ARTIFACT VALUES					
Mean of Square Root of Criterion Reliability	0.77459	--	0.77459	0.77459	0.77459
Mean of Square Root of Predictor Reliability	0.92195	--	0.92115	0.92115	0.92115
Mean Restricted SD (Range Restriction)	0.67	--	--	0.67	--
Mean of C[B]	--	--	0.70072	--	0.70070

Note: **Bolded** numbers represent statistics for Fisher z transformations. All values have been truncated (*not* rounded off) after the fifth decimal. [A]Unweighted statistic. [B]Range restriction attenuation factor.

To facilitate the reporting of results, the values presented in the table have been truncated (*not* rounded off) after the fifth decimal. As previously noted, the FORTRAN program reports statistics computed on both the raw data and Fisher z transformations. Because the other programs all use the raw data, discussion of the results of this program are limited to the raw data statistics only. The z statistics are, however, also reported in the table.

Summary Statistics. Identical summary statistics were obtained for the total number of subjects and rs. Identical sample-weighted rs were obtained for McDaniel (1986a) and Hunter and Schmidt (1990) programs; the FORTRAN and Huffcutt et al. (1993) programs differed from the other two after the fourth decimal. It is worth noting that the unweighted mean of .27605 differed from the weighted mean only after the third decimal. The standard deviation of the observed rs were identical for the Huffcutt et al., Hunter and Schmidt, and McDaniel programs. The FORTRAN program, again, differed from these at the fourth decimal.

Variance-Due-to-Sampling-Error. Identical variance-due-to-sampling-error values were obtained by the Hunter and Schmidt (1990) and McDaniel (1986a) programs. The Huffcutt et al. (1993) program differed from these other two at the fourth decimal. Similar results were obtained for the sampling error percent-variance-accounted-for values. The two programs that reported a residual variance statistic (Huffcutt et al. 1993; and McDaniel, 1986a) obtained almost identical values.

The final sampling error statistic reported is the chi square test for the homogeneity of sample correlations. This test determines whether the unexplained variance in the correlations is significantly greater than zero. This statistic was reported by only the McDaniel (1986a) and FORTRAN programs and the values differed by 1.34102.

13

Mean True r's. After correcting for all artifacts, the Hunter and Schmidt (1990) and McDaniel (1986a) programs reported identical mean true rs. The Huffcutt et al. (1993) program differed at the third decimal. The value obtained by all three programs was, however, higher than the true value. This finding is consistent with other Monte Carlo studies that have demonstrated that the validity generalization procedure tends to overcorrect for statistical artifacts (e.g., Pease & Switzer, 1988).

The percent-variance-due-to-all-artifacts values were different for all the programs. The range of this difference was 7.53192 (Min = 401.38798, Max = 408.9199). The other attenuation artifact corrections reported in Table 1 were also different across the programs even though the magnitude of these differences were, again, relatively small. The summary of mean artifact values reported by the programs were all identical.

IV. DISCUSSION

Summary

This study investigated whether the choice of meta-analysis software could influence the outcomes of a Monte Carlo validity generalization study. Thus, the current study was limited to a functional test of four commonly used Schmidt and Hunter validity generalization-based meta-analysis software programs which were compared in terms of the similarity of obtained results when used to analyze the same data. A strength of the present study was the availability of known "population" values which served as true scores to assess the accuracy of the programs. Specifically, a measure of accuracy was the deviation from a true mean corrected r of .53, with a variance of zero.

Meta-Analysis Statistics. The results of this investigation generally indicate that at least in terms of the four programs compared, the choice of software does not seem to make much of a difference. While there were some differences in values obtained from the different programs, these differences tended to be very small, typically occurring in the fourth and sometimes fifth decimal place, and did not influence the meta-analytic outcomes. For instance, a primary meta-analysis statistic is the mean corrected true r. After correcting for predictor and criterion unreliability and range restriction, the three programs that reported this statistic all obtained values that were fairly similar. These programs, however, tended to overcorrect for artifacts such that the corrected rs were larger than the true r. This finding of overcorrection, is consistent with other validity generalization Monte Carlo studies (Pease & Switzer, 1988).

Another important statistic, the mean sample-weighted r, was also fairly similar for all the programs compared here; and so were the sampling error percentage-of-variance-

15

accounted-for values. An interesting finding was that the Fisher z values reported by the FORTRAN program tended to be slightly higher than other comparable values. This is consistent with Hunter and Schmidt's (1990) comment that using the Fisher z transformation results in estimates of some validity generalization parameters that might be positively biased because the z transformation gives larger weights to large correlations than to small ones. But as they also note and was manifest here, in practice, this usually does not make much difference in the final meta-analysis outcome.

Inconsistencies in meta-analyses of the same topic have been partially attributed to the effect of judgment calls in the implementation process (e.g., Wanous et al., 1989). However, in terms of the programs compared in the present study, it would seem that the choice of which software or program to use is one decision in the implementation of a validity generalization study that does not have a major impact on the study's outcomes. While there were some differences in the values obtained, they tended to be small and not much different from those obtained from computing, for example, a t-test on SAS or SPSS on the same data. So it would seem that in the absence of any profound computational differences, the choice of meta-analysis software programs should be determined by the match between the needs of the researcher, the statistics provided by the program, the capabilities and features of the program, and the computing environment in which the researcher feels most proficient.

Program and Software Summary

The following discussion will highlight some additional features and capabilities of each of the programs. (A list of all the meta-analysis statistics furnished by each program is provided in the Appendix; this is intended to serve as a guideline to the potential user.)

16

The FORTRAN Program. As previously indicated, this program was designed to run on mainframe computers. It is likely that PC's with a FORTRAN compiler should be able to execute the program code, although the present study only used the program on a mainframe computer. One recommendation is that users of this program need to be facile with FORTRAN code to be able to locate and change fixed format data input and print specifications. Additionally, all of the available statistics are program defined. There are no user-selectable tests or subanalyses available. Also, unlike the other programs used in the present study, this program corrects for sampling error only. While this may appear to be a limitation, sampling error is in fact the most widely corrected for artifact (Steiner et al., 1991), and has been demonstrated to account for at least 85% of the explained variance due to all artifacts (Pearlman, Schmidt & Hunter, 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1981). Thus there are instances when this might be sufficient.

With the absence of any data auditing, error checking or diagnostic features (e.g., identifying r values greater than 1 or less than -1 in the dataset), data checking and diagnoses must be accomplished visually. Finally, no user's manual is available for this program.

Huffcutt, Arthur and Bennett's (1993) SAS PROC MEANS Program. This program uses the PROC MEANS procedure in SAS to compute most of the calculations and summary statistics called for in the Schmidt and Hunter meta-analysis approach. The code is transportable to any SAS operating environment, including PC-SAS. Like the Hunter and Schmidt (1990) and McDaniel (1986a) programs used in the current study, this program uses the artifact distribution approach. Artifact corrections are available and user-selectable for sampling error, range restriction, and unreliability. The program allows for the correction of any number and combination of artifacts. Further, potential moderator variables can be easily

17

coded and analyzed by simply including a few additional statements in the SAS program to sort and reanalyze the data according to the levels of the moderator variables. One limitation of this program is that even though relatively simple and easy, a few of the final calculations have to be made manually. Also, this program does not have a data auditing and error checking capability for either the developed artifact distributions or the primary dataset. Finally, a detailed user's manual is currently being developed; meanwhile, guidelines for users can be found in Huffcutt et al. (1993).

Hunter and Schmidt's (1990) VGNON.BAS BASIC Program. Of the four programs used in this paper, this is the only one that runs specifically on IBM-PC compatible microcomputers using GW-BASIC. However, mainframes with a resident BASIC compiler should be able to execute the program as well. Users of this program should note that the study N-size is preset to 100 studies. If the user has more than 100 studies, then the dimension array statement in the BASIC code, must be modified to accommodate the additional studies. The program uses the non-interactive approach and provides artifact corrections for sampling error, range restriction, and unreliability. The user must develop external artifact distribution data files for inclusion in the analyses. Thus input to the program consists of four data files. The format and structure of these files, which is fairly rigid and inflexible, is specified in Hunter and Schmidt (1990). In addition, all artifact corrections are program defined.

The analysis of moderators with this program requires that the user manually re-sort and separately re-enter the datasets. This will need to be done for each level of each moderator. It is also important to note that this program does not provide any data auditing and error checking capability for either the artifact distributions or the dataset. Finally,

information related to running of this program (and three other meta-analysis BASIC programs) is provided in Hunter and Schmidt (1990).

McDaniel's (1986a) MAME Program. This program consists of a series of SAS macros designed to calculate a variety of meta-analysis statistics which have been previously discussed. The present study used the MAME program on a mainframe computer running SAS version 5.18. Although mainframe SAS version 6.06 was available, the program code would not execute under the newer version. The code is also designed to run in PC-SAS version 6.03 and 6.04. The macro code contains several meta-analysis techniques which can be invoked by the user. Similar to the Hunter and Schmidt (1990) and Huffcutt et al. (1993) programs, artifact corrections are available for sampling error, range restriction, and unreliability and the user can select any or all of the corrections to be performed. The program also includes separate code to generate both default and user-defined artifact distributions. Further, the MAME program was the only program in the present study to offer data auditing and error checking capabilities for both the created artifact distribution and the dataset. These capabilities include (1) checking for r values that are either greater than 1 or less than -1; (2) identifying datasets with less than two observations; (3) identifying missing sample size, correlation and/or mean correlation values, d values, reliability coefficients, and frequencies in the dataset and artifact distributions; and (4) identifying reliability coefficients less than or equal to zero, and reliability values greater than 1.0. Finally, a detailed user manual is available from the program author.

In conclusion, it is suggested that future meta-analytic studies make the reporting of the specific data analysis programs and procedures used an integral part of their methodology. While the absence of major differences between the results generated by the different programs

was reassuring, the increasing reliance on validity generalization, differences in the statistics available, and the features and capabilities of these programs does warrant a call for researchers to report information regarding how the analyses in meta-analysis studies are performed.

REFERENCES

Arthur, W., Jr., Barrett, G V., & Alexander, R. A. (1991). Prediction of vehicular accident involvement: A meta-analysis. *Human Performance, 4*, 89-105.

Barrick, M. R. & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology, 65*, 543-558.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. New York: W, H, Freeman.

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry level jobs. *Journal of Applied Psychology, 79*, 184-190.

Huffcutt, A. I., Arthur, W., Jr., & Bennett, W. R., Jr. (1993). Conducting meta-analysis using the Proc Means procedure in SAS. *Educational and Psychological Measurement, 53*, 119-131.

Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper and I. T. Robertson (Eds.). *International Review of Industrial and Organizational Psychology 1987*. London, Gt. Britain: John Wiley.

Hunter, J. E., & Hunter R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury, CA: Sage.

Lord, R. G., DeVader, C. L., & Alliger, G. M. (1986). A meta-analysis of the relation between personality traits and leadership perceptions: An application of validity generalization procedures. *Journal of Applied Psychology, 71*, 402-410.

McDaniel, M. A. (1986a). *MAME: Meta-analysis made easy. Computer program and manual, ver 2.1.* Bethesda, MD: Author.

McDaniel, M. A. (1986b). Computer programs for calculating meta-analysis statistics. *Educational and Psychological Measurement, 64,* 175-177.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology, 73,* 327-330.

Pease, P. W., & Switzer, F. S., III. (1988). Validity generalization and hypothetical reliability distributions: A test of the Schmidt-Hunter procedure. *Journal of Applied Psychology, 73,* 267-274.

Rothstein, H. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322-327.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407-422.

Steiner, D. D., Lane, I. M., Dobbins, G. H., Schnur, A., & McConnell, S. (1991). A review of meta-analyses in organizational and human resources management: An empirical assessment. *Educational and Psychological Measurement, 51,* 609-626.

Wanous, J. P., Sullivan, S. H., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74,* 259-264.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analysis investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61,* 275-290.

Appendix:

Meta Analysis Statistics and Features of Programs Compared

Program: *FORTRAN Program*

<u>Meta-Analysis Statistics</u>	<u>Program Features</u>
Total N	• Transportable to any mainframe or PC with a FORTRAN compiler.
• No. r's	
• Obs SD	
• Pred SD	• User needs familiarity with FORTRAN code to change fixed format data input and print specifications.
• % Var Acct For	
• Residual SD	
• Mean r	
• Unbiased % Var	• All statistics are program defined. No user-definable capability.
• Chi Square	
• Crit Val .05	
• Confidence Intervals on: R, N-size	• Artifact correction only available for sampling error.
• Confidence Intervals on: R, N-size, & Resid Var	• No data auditing or error checking capability in the program.
	• No user's manual available.

Appendix - Contd.

Program: *Huffcutt, Arthur, and Bennett (1993)*

<table>
<tr><td><u>Meta-Analysis Statistics</u></td><td><u>Program Features</u></td></tr>
<tr><td>

- Total N
- Number of $r's$
- Mean Observed $r's$
  (sample-weighted)
- SD of Observed $r's$
- Variance Due to Sampling Error
- Percent Variance Accounted For
- Residual Variance
- Mean True r
- SD True r
- Variance Due to All Artifacts
- Residual Variance
- Mean of Square Root of
  Criterion Reliability
- Mean of Square Root of
  Predictor Reliability
- Mean Restricted SD
  (range restriction)

</td><td>

- Analysis based on SAS PROC MEANS procedures.

- Transportable to any SAS mainframe or PC-SAS environment.

- Some final calculations must be accomplished by hand.

- User-selected artifact corrections.

- Artifact corrections available for sampling error, range restriction, and unreliability.

- Moderator variable specification easily coded and analyzed.

- No data auditing or error checking capability provided.

- User manual currently under development.

</td></tr>
</table>

Appendix - Contd.

Program: *Hunter and Schmidt (1990)*

Meta-Analysis Statistics	Program Features
• Mean Observed *r* • *SD* of Observed *r's* • Predicted *SD* • % Var Acc for • Residual *SD* • Residual Var • Number of *r's* • Total *N* • Mean True Score *r* • *SD* of True Score *r* • Mean True Validity • *SD* of True Validity • Best Case • Worst Case	• Program will run on most IBM compatible personal computers, and possibly on mainframes with a BASIC compiler. • Program code written in BASIC. • Study *N*-size preset to 100 studies. If user has more than 100 studies, then BASIC code (dimension array statement) must be modified to accommodate this. • Artifact corrections are program defined. No user-specified corrections available. • User must develop artifact distributions for inclusion in analysis.

SUPPLEMENTARY RESULTS

• Total Variance • Sampling Error Var • % Var Due to Sampling Error • Var Due to Criterion Rel Diffs • Var Due to Test Rel Diffs • Var Due to Range Res Diffs • Mean of SQR of Criterion Rel • Mean of SQR of Test Rel • Mean Restricted *SD* • Mean *R* Corrected for Range Res	• Artifact corrections available for sampling error, range restriction, and unreliability. • To analyze for potential moderator variables, data must be manually sorted and separately re-entered, i.e., different data set must be created for each level of moderator variable. • No data auditing or error checking capability provided. • User information available in Hunter & Schmidt (1990) book.

Appendix - Contd.

Program: *McDaniel (1986a)*

<u>Meta-Analysis Statistics</u>

<u>Program Features</u>

DESCRIPTIVE STATISTICS

- Number of Correlation Coefficients
- Total Number of Observations
- Mean Observed Correlation
- Observed Variance
- Observed *SD*

SAMPLING ERROR RESULTS

- Variance Due to Sampling Error
- (SE Variance based on Mean *r*)
- *SD* Due to Sampling Error
- Percent Variance Accounted For
- Residual Variance
- Residual Standard Deviation
- Reliability of Correlation Vector
- Chi Square Test for Significance
- Significance of Chi Square

ARTIFACT VARIANCE RESULTS

- Var. Due to Unrel. & Range Restriction
- Variance Due to All Artifacts
- *SD* Due to All Artifacts
- *SD* Due to All Artifacts
- Percent of Variance Due to All Artifacts
- Residual Variance
- Residual *SD*

- Program based on a series of SAS Macros. Transportable to any mainframe running SAS and PCs running PC-SAS version 6.03 or 6.04.

- Current version of Macro code is incompatible with mainframe SAS version 6.06. Will run on version 5.18 (or earlier).

- Program also includes code to calculate default artifact distributions for use in analyses.

- Same distribution code can be used to generate user-specified artifact distributions.

- Macro code contains several separate meta-analysis techniques.

- Macro code allows for user-specified artifact corrections.

- Artifact corrections available for sampling error, range restriction, and unreliability.

- Comprehensive data auditing and error correction capability provided for artifact distribution creation and data set development.

- Detailed user's manual available.

Appendix - Contd.

Program: *McDaniel (1986a)* - Contd.

Meta-Analysis Statistics	Program Features
MEAN CORRECTED FOR	
A AND C[1]	
VARIANCE CORRECTED	
FOR A, B, and C	
• Mean R	
• True Variance	
• True SD	
• Bottom 10th Percentile	
MEAN AND VARIANCE	
CORRECTED FOR A, B, and C	
• Mean r	
• True Variance	
• True SD	
• Bottom 10th Percentile	
MEAN ARTIFACT VALUES	
• Mean of A	
• Mean of B	
• Mean of C	

Note: [1]A = criterion reliability; B = predictor reliability; C = range restriction.